**1. Deriving Regression Estimators by Hand**

* This involves understanding the mathematical foundation of regression.
* The least squares method is used to derive the estimators (slope and intercept) that minimize the sum of squared residuals.
* You'll work through the derivation of:
  + β^1=∑(xi−xˉ)(yi−yˉ)/∑(xi−xˉ)2 (Slope)
  + β^0=yˉ−β^x (Intercept)

**2. Fitting Models Using R and Interpreting Outputs**

* You'll learn how to fit a linear regression model in R using lm() function.
* The output includes:
  + **Coefficients** (slope and intercept)
  + **R-squared value** (explains variance explained by the model)
  + **p-values** (assess significance of predictors)
  + **Residual standard error** (indicates model accuracy)
* Interpretation involves determining how well the model describes biological relationships.

**3. Examining How Environmental Pressures Shape Trait Relationships**

* Regression can be used to assess how environmental variables (e.g., temperature, habitat, pollution) influence biological traits (e.g., size, fitness, gene expression).
* This can be linked to ecological and evolutionary studies to infer adaptation.

**4. Linking Statistical Outputs to Evolutionary Hypotheses**

* Statistical models help test hypotheses about trait evolution.
* Example: A strong correlation between temperature and body size may support **Bergmann’s rule** (larger body size in colder climates).
* p-values and confidence intervals help determine if observed relationships are likely due to evolutionary selection.

**5. Evaluating Model Assumptions Using Diagnostic Tools**

* **Linearity:** Checking scatter plots and residual plots.
* **Independence:** Ensuring observations are not correlated (e.g., using Durbin-Watson test).
* **Normality:** Checking residuals with histograms or QQ plots.
* **Homoscedasticity:** Ensuring constant variance in residuals (e.g., using Breusch-Pagan test).
* **Collinearity:** Checking Variance Inflation Factor (VIF) for predictor correlations.

**1. Deriving Regression Estimators by Hand**

Linear regression models the relationship between an **independent variable** (predictor, X) and a **dependent variable** (response, Y) using the equation:

Y=β0+β1X+ϵ

where:

* Y is the dependent variable.
* X is the independent variable.
* β0 (Intercept) is the predicted value of Y when X=0.
* β1​ (Slope) is the change in Y per unit change in X.
* ϵ is the residual error (unexplained variation).

**Mathematical Derivation**

The best-fitting regression line minimizes the sum of squared residuals:

SSE=∑(Yi−(β0+β1Xi))2

Taking derivatives and solving for β0​ and β1​, we get:

β1=∑(Xi−Xˉ) (Yi−Yˉ)/∑(Xi−Xˉ)2

β0​=Yˉ−β1​Xˉ

**Multiple regression, Polynomial regression and logistic regression**

 **Multiple Regression:** More than one predictor (e.g., temperature, pH, nutrients).

 **Polynomial Regression:** Used for curved relationships (e.g., enzyme activity vs. substrate).

 **Logistic Regression:** Used when the outcome is binary (e.g., drug response).

**Example with iris dataset in R**

| **Questions** | **Best Statistical Test** |
| --- | --- |
| Relationship between **Sepal.Length and Sepal.Width** | **Linear Regression** |
| Impact of **multiple features on Sepal.Length** | **Multiple Regression** |
| Non-linear relationship between **Sepal.Length and Sepal.Width** | **Polynomial Regression** |
| Predicting **Species from Sepal & Petal features** | **Logistic Regression** |
| Testing if **Sepal.Length differs between Species** | **ANOVA** |

**Key Differences Between Regression Types and ANOVA**

| **Method** | **Type of Relationship** | **Response Variable Type** | **Predictors** | **Example Use Case** |
| --- | --- | --- | --- | --- |
| **Linear Regression** | Linear | Continuous | One predictor (X) | Height vs. Weight |
| **Multiple Regression** | Linear | Continuous | Multiple predictors (X1, X2, X3...) | Gene expression vs. Drug dose, Time |
| **Polynomial Regression** | Non-linear | Continuous | One or more predictors (with polynomial terms) | Growth rate vs. Temperature |
| **Logistic Regression** | Logistic (S-curve) | Categorical (Binary/Multiclass) | One or more predictors | Disease presence (Yes/No) |
| **ANOVA** | Group comparisons | Continuous | Categorical predictors (factor levels) | Gene expression across different treatments |

**Choosing the Right Model (When to Use What?)**

* **Linear Regression**: When **one continuous predictor** influences a **continuous response**.
* **Multiple Regression**: When **multiple predictors** affect a **continuous response**.
* **Polynomial Regression**: When the relationship between predictors and response is **non-linear**.
* **Logistic Regression**: When predicting a **binary or categorical outcome**.
* **ANOVA**: When comparing **group means** (e.g., gene expression across **multiple drug treatments**).

**Notes**

1. Variance measures how spread-out X values are, while covariance shows the direction and strength of the linear relationship.

S2X​= ∑(Xi​−Xˉ)2/n-1, in r programming , var (x)

SXY​=∑(Xi​−Xˉ)(Yi​−Yˉ)​/n-1, in r programming, cov(x, y)

1. Residuals (ei) measure deviations from the fitted line.

**How does SE(β1​) relate to SE(xˉ) ?**

* **SE(β1​) measures uncertainty in the slope estimate**, while **SE(xˉ) quantifies the uncertainty in the mean**. Both depend on sample variance, but:
  + **SE(β1) is influenced by the spread of X-values** (higher variance in X leads to a more precise slope estimate).
  + **SE(xˉ) decreases as sample size increases**, providing a more accurate estimate of the mean.
  + A wider range of X-values reduces **collinearity**, leading to a lower **SE(β1​)** and a more reliable estimate of metabolic scaling